How to Analyze Paired Comparison Data

Kristi Tsukida and Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
{gupta}@ee.washington.edu

UWEE Technical Report
Number UWEETR-2011-0004
May 2011

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
MAY 2011

2. REPORT TYPE

3. DATES COVERED
00-00-2011 to 00-00-2011

4. TITLE AND SUBTITLE
How to Analyze Paired Comparison Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington,Department of Electrical Engineering,Seattle,WA,98195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Thurstone’s Law of Comparative Judgment provides a method to convert subjective paired compar- isons into one-dimensional quality scores. Applications include judging quality of different image reconstructions, or different products, or different web search results, etc. This tutorial covers the popular Thurstone-Mosteller Case V model and the Bradley-Terry logistic variant. We describe three approaches to model- fitting: standard least-squares, maximum likelihood, and Bayesian approaches. This tutorial assumes basic knowledge of random variables and probability distributions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**19**	

Abstract

Thurstone's Law of Comparative Judgment provides a method to convert subjective paired comparisons into one-dimensional quality scores. Applications include judging quality of different image reconstructions, or different products, or different web search results, etc. This tutorial covers the popular Thurstone-Mosteller Case V model and the Bradley-Terry logistic variant. We describe three approaches to model-fitting: standard least-squares, maximum likelihood, and Bayesian approaches. This tutorial assumes basic knowledge of random variables and probability distributions.

Contents

1 Why Paired Comparisons?

When comparing different options, one often wishes to assign a single quality score to each option. For example, you may want to score the quality of different image processing algorithms, or the skill of different chess players, or the seriousness of different crimes. To illustrate, we place three images of an apple on a quality scale in Figure 1. Here, we are concerned with scores on an *interval scale*, which means that only the relative differences between scores matters, and the scores could be shifted or multiplicatively scaled without changing their meaning. An interval scale has two degrees of freedom: the placement of zero, and the unit of measurement. Fahrenheit and Celcius temperature scales are interval scales, because the difference between two temperatures matters, but the placement of 0° and the unit of "1 degree" is arbitrary.[1]

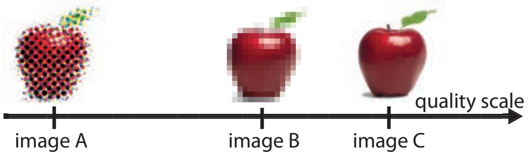

Figure 1: Image quality scale

How do you gather data to score a set of options? It is tempting to simply ask a bunch of people to score each of your options: "On a scale of 1 to 10, what is the quality of this image?" However, people may mean different things by the same score (a 3 to one person may be different than a 3 for another person). It may be hard to determine specifically what "1" and "10" mean (how bad does an image have to be to get a 1?). It may be inflexible (what if you want to give something a 15?). Further, you may care more about scoring the options in the context of the set, rather than on an absolute scale (you want to know "How much better does this image look than the other options?" rather than "Does this image look good?"). Because of these issues, gathering paired comparisons may be more useful than directly asking for quality scores.

In a paired comparison experiment, you ask, "Is A better than B?" Generally ties are not allowed (or they may be counted as half a vote for each option). Ideally you would get comparisons for all possible pairs of options you are judging, but this is not necessary to estimate the scores, and for a large number of options, may simply be infeasible. There may also be issues of order presentation (which option is presented first could affect the preference) but in the rest of this tutorial we assume that this issue can be ignored.

The result of a paired comparison experiment is a count matrix, C, of the number of times that each option was preferred over each other option.

$$C_{i,j} = \begin{cases} \text{\# times option } i \text{ preferred over option } j & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

1.1 Roadmap

Now that we have defined the problem and the experimental count data, in the next Section we describe Thurstone's statistical model of judgments and Thurstone's Law of Comparative Judgment, which provide a method of estimating the quality score difference for two options. We will then extend the analysis to estimating scores for more than two options using a least squares method (Section 3), a maximum likelihood method (Section 4), and using the expected value of the score (Section 5). We illustrate the different approaches with simulations (Section 6). This document ends with Matlab code to implement the described functions.

[1] An alternate is a *ratio scale*, where the zero value is fixed e.g. age, where 0 years is a specific reference point. In a ratio scale we can always compare to 0, so 20 years is twice as old as 10 years, whereas on a interval scale 20° is not twice as hot as 10°. The classification of measurement scales is discussed by Stevens [10].

2 Models for Comparative Judgment

There are two common models for analyzing paired comparison data. We first discuss Thurstone's model, and then the Bradley-Terry model.

2.1 Thurstone's Model

In 1927, Louis Leon Thurstone pioneered psychometrics by using Gaussian distributions to analyze paired comparisons [11, 7]. Thurstone's model assumes that an option's quality is a Gaussian random variable.

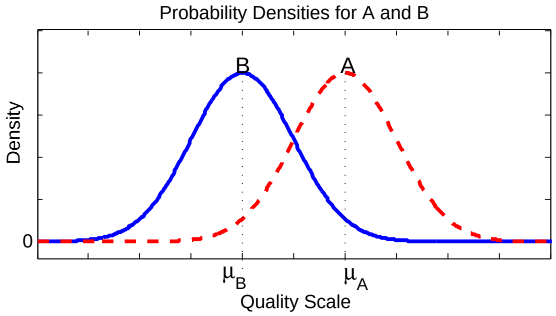

Figure 2: PDFs of A and B

Consider the basic case of two options, where we let the Gaussian random variables A and B represent the quality of option A and option B respectively:

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2), \qquad B \sim \mathcal{N}(\mu_B, \sigma_B^2),$$

$$P(A = a) = \tfrac{1}{\sigma_A}\,\phi\left(\tfrac{a-\mu_A}{\sigma_A}\right), \qquad P(B = b) = \tfrac{1}{\sigma_B}\,\phi\left(\tfrac{b-\mu_B}{\sigma_B}\right),$$

where ϕ is the standard normal PDF (zero mean, unit variance),

$$\phi(x) = \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

Thurstone's model says that when a person judges whether A is better than B, they draw a realization from A's quality distribution and a realization from B's quality distribution, and then choose the option with the higher quality. Equivalently, they choose option A over option B if their draw from the random variable $A - B$ is greater than zero,

$$P(A > B) = P(A - B > 0).$$

Since $A - B$ is the difference of two Gaussians, $A - B$ is a Gaussian random variable:

$$\begin{aligned} A - B &\sim \mathcal{N}(\mu, \sigma) \\ \mu_{AB} &= \mu_A - \mu_B \\ \sigma_{AB}^2 &= \sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B, \end{aligned} \tag{1}$$

where ρ_{AB} is the correlation between A and B.

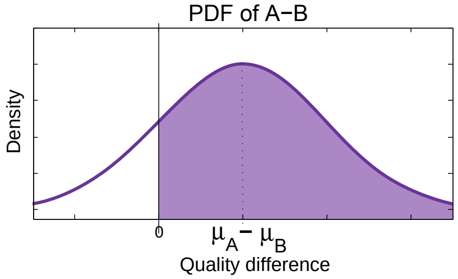

Figure 3: $P(A > B)$ is the shaded area under the PDF of $A - B$.

Therefore the probability of choosing option A over option B (shown in Figure 3) is

$$\begin{aligned} P(A > B) &= P(A - B > 0) \\ &= \int_0^\infty \tfrac{1}{\sqrt{2\pi\sigma_{AB}^2}}\, e^{-(x-\mu_{AB})^2/(2\sigma_{AB}^2)} dx \\ &= \int_{-\mu_{AB}}^\infty \tfrac{1}{\sqrt{2\pi\sigma_{AB}^2}}\, e^{-x^2/(2\sigma_{AB}^2)} dx \\ &= \int_{-\infty}^{\mu_{AB}} \tfrac{1}{\sqrt{2\pi\sigma_{AB}^2}}\, e^{-x^2/(2\sigma_{AB}^2)} dx \\ &= \int_{-\infty}^{\mu_{AB}} \tfrac{1}{\sigma_{AB}}\, \phi(\tfrac{x}{\sigma_{AB}}) dx \\ &= \int_{-\infty}^{\frac{\mu_{AB}}{\sigma_{AB}}} \phi(t) dt \\ &= \Phi\left(\frac{\mu_{AB}}{\sigma_{AB}}\right), \end{aligned} \tag{2}$$

where $\Phi(z)$ is the standard normal cumulative distribution function (CDF)

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt = \int_{-\infty}^{z} \phi(t)\, dt.$$

By inverting (2), we can calculate the quality difference μ_{AB} as

$$\mu_{AB} = \sigma_{AB}\, \Phi^{-1}\left(P(A > B)\right),$$

where $\Phi^{-1}(x)$ is the inverse CDF of the standard normal. The inverse CDF of the standard normal is also commonly known as the *z-score* or *standard score* since it gives the number of standard deviations that x is from the mean. Although traditionally getting the z-score required lookup tables, modern computers can calculate the inverse CDF function precisely.

Thurstone proposed estimating $P(A > B)$ by the empirical proportion of people preferring A over B, $C_{A,B}/(C_{A,B} + C_{B,A})$. So, the estimator for the quality difference $\hat{\mu}_{AB}$ is

$$\hat{\mu}_{AB} = \sigma_{AB}\, \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right). \tag{3}$$

$$\tag{4}$$

The estimate (3) is known as Thurstone's Law of Comparative Judgment.

2.2 Thurstone's Case V Model

Thurstone made a number of model simplifications for tractability. The estimate (3) requires σ_{AB} to be known or estimated, and in the general model that requires estimating the variances and correlation of A and B as per (1). The simplest and most popular simplification is the Case V model, which assumes that all options have equal variance and zero correlations (or less restrictively, equal correlations instead of zero correlations [9]):

$$\sigma_A^2 = \sigma_B^2$$
$$\rho_{AB} = 0.$$

Without loss of generality[2], set the variances to one half $\sigma_A^2 = \sigma_B^2 = \frac{1}{2}$ so the variance of $A - B$ is one,

$$\sigma_{AB}^2 = \sigma_A^2 + \sigma_B^2 = 1.$$

This sets the scale unit for the interval scale. This simplifies Thurstone's Law given in (3) for Case V to

$$\hat{\mu}_{AB} = \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right). \tag{5}$$

For the rest of this tutorial, we will use these Case V assumptions and refer to this Case V model as *Thurstone's model*.

Another common approach is to assume that $\sigma_A^2 = \sigma_B^2 = 1$, so that $\sigma = \sqrt{2}$, and Thurstone's law for Case V carries an extra constant of $\mu = \sqrt{2}\Phi^{-1}(\frac{C_{A,B}}{C_{A,B}+C_{B,A}})$. This makes the equations less convenient, but then the quality scale differences are easily interpreted as the number of standard deviations. If you want the quality scaled in terms of standard deviations, you can multiply the quality scale values in this tutorial by $\sqrt{2}$ (which is ok since the interval scale is not affected by multiplicatively scaling the scores).

2.3 Prior Knowledge

Prior knowledge is easily incorporated into the model by adding values to the count matrix according to what you believe the proportion of counts should be *a priori*. Create a matrix B of the proportion of times you believe one option would be preferred over the other. Then add a weighted version to the collected data.

$$\tilde{C} = C + \alpha B.$$

Using a prior B can help regularize the estimates and solve the 0-1 problem, as we discuss in Section 3.2.

2.4 The Bradley-Terry model

Bradley and Terry [2] introduced an alternate model, also known as the Bradley-Terry-Luce model (BTL) for Luce's extension to multiple variables in [6]. The BTL model differs from the Thurstone model in that it uses Gumbel random variables for the quality of each option instead of a Gaussian. Then the BTL scale difference $A - B$ is a logistic random variable, and so $P(A > B)$ can be calculated from the logistic cumulative distribution function. This model has a closed-form solution:

$$P(A > B) = P(A - B > 0) = \frac{\exp(\mu_A/s)}{\exp(\mu_A/s) + \exp(\mu_B/s)} \tag{6}$$
$$= \frac{1}{2} + \frac{1}{2}\tanh(\frac{\mu_A - \mu_B}{2s}), \tag{7}$$

where s is the scale parameter for the logistic distribution (changing s changes the variance, but not the mean). Equation (6) says that the probability. Then by again estimating $P(A > B)$ by the empirical count proportion $\frac{C_{A,B}}{C_{A,B}+C_{B,A}}$ and inverting (7) yields the estimate

$$\hat{\mu}_A - \hat{\mu}_B = s\left(\ln\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right) - \ln\left(1 - \frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right)\right). \tag{8}$$

The inverse logistic CDF in (8) is commonly called the *logit*. To compare the BTL model scale differences with the Thurstone model ones from (3), equate the variance by setting $s = \frac{\sqrt{3}}{\pi}$. Empirically as shown in Figure 4, the logistic CDF is very similar to the Gaussian CDF, so that using Thurstone-Mosteller's solution or BTL's model produces very similar results. Some people prefer BTL for computational simplicity (you don't have to compute the inverse Gaussian CDF), although with modern computers and algorithms, computing the inverse Gaussian CDF is simple, so the computational aspect is not an issue.

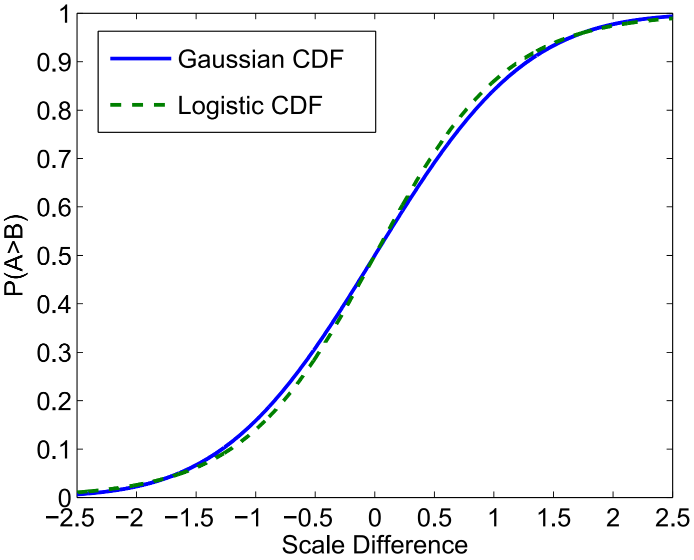

Figure 4: Gaussian vs Logistic CDF

The logistic CDF has a fatter tail and is slightly more sloped at the inflection point than a Gaussian with the same mean and variance. This means that the BTL model will estimate slightly smaller scale differences for proportions near $\frac{1}{2}$ and slightly larger scale differences for proportions near 0 or 1 when compared with Thurstone's model.

3 Model Fitting

Thurstone's model provides a method of estimating the scale difference for any single pair of options by estimating $P(A > B)$ by the empirical proportion of people preferring A to B. However, when considering more than two options this approach breaks down because these values need to be massaged to fit on a one dimensional scale. In this section we detail three different approaches to estimating the quality scores given more than two options for the Thurstone model (the same approaches can be applied to the BTL model).

3.1 Thurstone-Mosteller Least Squares Method

To determine the quality scores for a set of m options, Thurstone offered a solution, which Mosteller later showed was the solution to a least squares optimization problem [9]. Define the vector of quality scores $\mu = [\mu_1, \mu_2, ..., \mu_m]$, and let D be an $m \times m$ matrix where $D_{i,j} = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j}+C_{j,i}}\right)$ is the (Case V) Thurstone's Law estimate (5) for the quality difference between option i and option j. (You may also use the BTL model by forming D using the logit from (8).) The least squares estimate for the quality scores μ minimizes the squared error between the quality scores and the Thurstone's Law pairwise estimates:

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^m} \sum_{i,j} (D_{i,j} - (\mu_i - \mu_j))^2. \tag{9}$$

This least squares problem has a simple closed-form solution which can be derived from the D matrix. If we set $\mu_1 = 0$, the least squares solution is

$$\hat{\mu}_j = \sum_{i=1}^{m} \frac{D_{i,1}}{m} - \sum_{i=1}^{m} \frac{D_{i,j}}{m}. \tag{10}$$

If instead of assuming $\mu_1 = 0$, you prefer to assume that the mean of all the μ_i is zero, the solution is

$$\mu_j = \sum_{i=1}^{m} \frac{D_{i,j}}{m}. \tag{11}$$

3.2 The 0/1 Problem

When estimating quality differences by this least squares method, we will have problems estimating when the proportion $\frac{C_{A,B}}{C_{A,B}+C_{B,A}}$ is 0 or 1, since $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$.

There are a couple of standard ways to deal with this problem [3], and we note a third optimal solution. One solution to the 0/1 problem is to simply ignore the 0/1 entries and use an incomplete matrix solution [8, 4]. We argue this is too heavy-handed a fix in that it ignores important information that the one option is strongly preferred to the other option.

Another standard solution is to "fix" the 0/1 proportions by modifying the entries in the count data matrix to change the 0/1 proportions to be $\frac{1}{n}$ and $1 - \frac{1}{n}$, where n is the number of people surveyed for that comparison. This is equivalent to modifying the count data:

$$\tilde{C}_{ij} = \begin{cases} \frac{1}{2} & \text{if } C_{ij} = 0 \\ C_{ij} - \frac{1}{2} & \text{if } C_{ij} = n \text{ (or equivalently if } C_{ji} = 0) \\ C_{ij} & \text{otherwise} \end{cases} \tag{12}$$

We will refer to this modified data matrix as the *0/1 fixed data*. A related solution is to add a count or fractional count to both the 0 and 1 entries of the count matrix; this is equivalent to assuming some prior data (see Section 2.3). These fixes do change the count matrix, but in a conservative way that biases the counts toward less confidence, and this fix is not as big a change as simply ignoring 0/1 entries. If one employs this fix, we recommend adding 1/2 count to adequately fix the 0/1 problem without adding too much noise to the data, though we note there is no general optimal value.

We note a third solution if there are more than two options: estimate the means by the maximum likelihood estimate, which we detail in Section 4.2. This is the solution we recommend because it is an optimal approach to the estimation and does not require adding noise or ignoring data.

4 Maximum Likelihood Scale Values

First, we show that if there are just two options, the maximum likelihood estimate of the quality score difference is given by Thurstone's law (3). Then we give the maximum likelihood estimate for multiple options (which is not equivalent to the least-squares solution given above).

4.1 Maximum Likelihood for Two Options

Let the paired judgment data for two options, A and B, be $a = C_{A,B}$ and $b = C_{B,A}$. The likelihood of the quality difference, $\mu_{AB} = \mu_A - \mu_B$, is

$$\begin{aligned} L(\mu_{AB}) = P(a, b|\mu_{AB}) &= \frac{1}{\gamma} P(A > B)^a P(B > A)^b \\ &= \frac{1}{\gamma} \Phi(\mu_{AB})^a \Phi(-\mu_{AB})^b, \end{aligned} \tag{13}$$

where γ is a normalization constant and we have used the identity $P(B > A) = 1 - \Phi(\mu_{AB}) = \Phi(-\mu_{AB})$. The maximum likelihood quality difference is

$$\begin{aligned} \hat{\mu}_{AB} &= \arg\max_{\mu_{AB}} L(\mu_{AB}) \\ &= \arg\max_{\mu_{AB}} \left(\Phi(\mu_{AB})\right)^a \left(\Phi(-\mu_{AB})\right)^b \\ &= \arg\max_{\mu_{AB}} a \log\left(\Phi(\mu_{AB})\right) + b \log\left(\Phi(-\mu_{AB})\right). \end{aligned}$$

This may be solved by Lagrange multipliers

$$\begin{aligned} 0 &= \frac{a}{\Phi(\mu_{AB})} \phi(\mu_{AB}) - \frac{b}{\Phi(-\mu_{AB})} \phi(-\mu_{AB}) \\ \frac{a}{\Phi(\mu_{AB})} &= \frac{b}{1 - \Phi(\mu_{AB})} \\ \hat{\mu}_{AB} = \Phi^{-1}\left(\frac{a}{a+b}\right) &= \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right), \end{aligned}$$

which verifies that Thurstone's Law yields the maximum likelihood solution if there are only $m = 2$ options.

4.2 Maximum Likelihood for Multiple Options

Extending the two option maximum likelihood estimation described in Section 4, to a comparison of m options, there is no longer a closed-form solution. Instead, one must solve a optimization problem.

Let μ be the vector of quality scale values $\mu = [\mu_1, \mu_2, ..., \mu_m]$. Define the log-likelihood of our count data, C, as

$$\mathcal{L}(\mu|C) \triangleq \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)). \tag{14}$$

To find the maximum likelihood solution quality scale values, one must solve

$$\begin{aligned} \arg\max_{\mu} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) \\ \text{subject to} \quad & \sum_i \mu_i = 0. \end{aligned} \tag{15}$$

To find a unique solution, include the constraint that the mean of all the quality scale values is zero as in (15), or set one of the quality scale values to zero $\mu_1 = 0$.

We show in the appendix that (15) is a convex optimization problem.

4.3 Maximum A Posteriori Estimation

One can also form the maximum a posteriori (MAP) estimate, by including a prior on the scale values $p(\mu)$:

$$\begin{aligned} \arg\max_{\mu} \quad & \mathcal{L}(\mu|C) + \log(p(\mu)) \\ \text{subject to} \quad & \sum_i \mu_i = 0. \end{aligned}$$

If there is little information about the true scale values that can be used to choose a prior, then we suggest a Gaussian prior that assumes the different scale values are drawn independently and identically from a standard normal will reduce the estimation variance and often provide better estimates. In that case the MAP estimate solves:

$$\begin{aligned} \arg\max_{\mu} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) - \sum_i \frac{\mu_i^2}{2} \\ \text{subject to} \quad & \sum_i \mu_i = 0. \end{aligned} \tag{16}$$

This choice of prior acts as a ridge regularization on the scale values [5], and is still a convex optimization.

4.4 Advantages of Maximum Likelihood Estimation

Maximum likelihood estimation is an optimal approach to estimation problems in the sense that it produces the solution that is most likely. Additionally, this maximum likelihood solution does not suffer from the 0/1 problem of the least squares methods because the maximum likelihood method does not use the inverse CDF. The 0 entries in the count matrix do not contribute to the likelihood and μ_i are constrained by the other terms in the log-likelihood to keep them from being driven to ∞.

5 Expected Quality Scale Difference

Instead of using the maximum likelihood quality difference, one can estimate the quality difference to be the *expected quality difference* where the expectation is taken with respect to the likelihood (or with respect to the posterior mean if there is a prior). On average, we expect this solution to perform better since this approach uses the full likelihood information rather than just the maximum of the likelihood.

5.1 Expected Quality Estimate

Consider the two option case. Treat the unknown quality score difference as a random variable U. The likelihood $P(a, b|U = u)$ is the probability of observing a people preferring option A, and b people preferring option B, as given in (13).

$$P(a, b|U = u) = \binom{a+b}{a} P(A > B|U = u)^a \, P(B > A|U = u)^b.$$

Using Bayes rule, the posterior can be written in terms of the likelihood as

$$P(U = u|a, b) = \frac{P(a, b|U = u)P(U = u)}{P(a, b)}.$$

$$\begin{aligned} P(U = u|a, b) &= \frac{1}{\gamma} P(A > B|U = u)^a \, P(B > A|U = u)^b \, P(U = u) \\ &= \frac{1}{\gamma} \Phi(u)^a \, (1 - \Phi(u))^b \, P(U = u), \end{aligned}$$

where γ is a normalizer constant,[3]

$$\gamma = \int_{\infty}^{\infty} \Phi(u)^a \left(1 - \Phi(u)\right)^b P(U = u) du.$$

If we assume a uniform prior for $P(U = u)$ over the range $[-t, t]$, then the expected quality scale difference U is

$$E[U|a,b] = \frac{1}{2t\gamma} \int_{-t}^{t} u\, \Phi(u)^a (1 - \Phi(u))^b \, du. \tag{17}$$

Alternatively, we could assume a Gaussian prior for $P(U = u)$. Using the standard normal, the expected quality scale difference U is

$$\begin{aligned} E[U|a,b] &= \frac{1}{\gamma} \int_{-\infty}^{\infty} u\, \Phi(u)^a (1 - \Phi(u))^b \, \phi(u) du \qquad (18) \\ &= \frac{1}{\gamma} \int_{0}^{1} \Phi^{-1}(p)\, p^a (1 - p)^b \, dp. \end{aligned}$$

5.2 Computation of Expected Quality Estimate

Unfortunately, no closed form solution exists, so the integral and the normalizer constant must be numerically computed. Numerical integration is slow, and may be prone to precision errors depending on the method of integration. Matlab may be used to attempt to approximate the integral (trapz, quad, quadgk), but Matlab is limited to machine precision (32 or 64 bits) and may not evaluate the integral accurately enough. (Matlab may return 0 when the actual solution should be a very small non-zero number). Mathematica has more sophisticated numerical integration routines and arbitrary precision calculations. Maple may also be used.

5.3 Bayesian Estimation of Preference Probability

In the previous section, we considered the quality difference, U, to be a random variable, and estimated it by taking its expectation. In this section, we instead consider the probability that option i is chosen over option j to be a random variable, $X_{i,j}$. We consider the result of performing Bayesian estimation, and the relation to priors and smoothing.

The count data is generated from a binomial distribution where the parameter $x_{i,j}$, is an observation of $X_{i,j}$:

$$\begin{aligned} P(C_{i,j}, C_{j,i} | x_{i,j}) &\sim \text{Binom}(C_{i,j} + C_{j,i} \,|\, x_{i,j}) \\ &\propto {x_{i,j}}^{C_{i,j}} (1 - x_{i,j})^{C_{j,i}} \end{aligned}$$

After observing the data C and assuming a uniform prior probability on $x_{i,j}$, the posterior probability of $x_{i,j}$ has a beta distribution with parameters $C + 1$:

$$\begin{aligned} P(x_{i,j} | C_{i,j}, C_{j,i}) &\propto {x_{i,j}}^{C_{i,j}} (1 - x_{i,j})^{C_{j,i}} \\ &\sim \text{Beta}(x_{i,j} | C_{i,j} + 1, C_{j,i} + 1). \qquad (19) \end{aligned}$$

The maximum a posteriori estimate of $x_{i,j}$ is $\hat{x}_{i,j} = \dfrac{C_{i,j}}{C_{i,j} + C_{j,i}}$ (the mode of the beta distribution (19)). Then calculate the quality scale difference by setting $\hat{x}_{i,j} = \Phi(u_i - u_j)$ and inverting, which results in Thurstone's law:

$$u_i - u_j = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j} + C_{j,i}}\right).$$

[3] Although at first glance $P(u|a,b)$ looks similar to a beta distribution, it is parameterized by u, instead of $p = \Phi(u)$, so it is not the same. The normalizer must be calculated numerically.

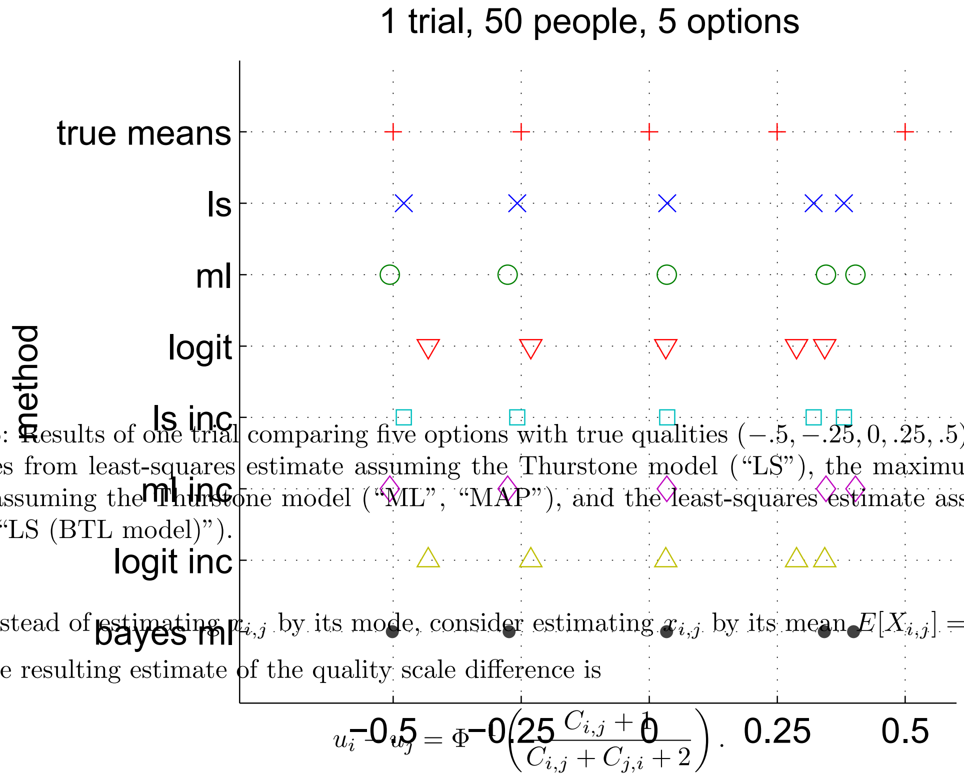

Figure 5: Results of one trial comparing five options with true qualities $(-.5, -.25, 0, .25, .5)$. Plot compares estimates from least-squares estimate assuming the Thurstone model ("LS"), the maximum likelihood estimate assuming the Thurstone model ("ML", "MAP"), and the least-squares estimate assuming the BTL model ("LS (BTL model)").

Next, instead of estimating $x_{i,j}$ by its mode, consider estimating $x_{i,j}$ by its mean $E[X_{i,j}] = \dfrac{C_{i,j}+1}{C_{i,j}+C_{j,i}+2}$. Then the resulting estimate of the quality scale difference is

$$u_i - u_j = \Phi^{-1}\left(\frac{C_{i,j}+1}{C_{i,j}+C_{j,i}+2}\right).$$

This result is equivalent to the Thurstone's law estimate if one puts a prior of 1 on all the counts, meaning that *a priori* you believe that all of the choices are possible. This may also be interpreted as Laplace smoothing the count data.

6 Illustrative Experiments

We illustrate the different estimation approaches with some simulations. We make n quality observations for each of pair of m options (simulating surveying n people for their preferences about all possible pairs of m options). We first generate the true means $\mu_1, \mu_2, ..., \mu_m$ for the m quality score distributions. Then, for each pairwise comparison for each person, the perceived quality scores for the ith option are drawn IID from $\mathcal{N}(\mu_i, \sigma^2 = \frac{1}{2})$ as in Thurstone's Case V. The count data is collected, and $\hat{\mu}$ is estimated from the data.

Fig. 5 illustrates the results of one run of the simulation, with $n = 50$ people surveyed about all pairs of $m = 5$ options. The true mean values are marked on top, and are at $(-.5, -.25, 0, .25, .5)$. The mean quality estimates are shown on the next rows for the least-squares estimate assuming the Thurstone model, the maximum likelihood estimate assuming the Thurstone model, and the least-squares estimate assuming the BTL model.

Fig. 6 shows results for a single pair of options ($m = 2$), averaged over 10,000 runs of the simulation for varying true quality-differences. If the true quality difference (shown on the x-axis) is large, then the 0/1 problem occurs (see Sec. 3.2), and the Thurstone's Law estimate given by (3) is that the quality difference is "infinite." For this case of $m = 2$ options, recall that (3) is also the maximum likelihood estimate. The green line is the maximum likelihood estimate given a prior of 1 count to both options (this could also be called a maximum a posterior estimate). This is always well-defined and performs better than Thurstone's Law for all quality differences. The red line shows the mean quality estimate as given by (17). This is a more robust estimate, and as shown in Fig. 6, will perform better than the maximum likelihood with prior when the true quality difference is large, but may perform worse when the true quality difference is small.

Fig. 7 shows the simulation results when there are $m = 10$ options. These results were averaged over 1000 runs of the simulation. For each run, the true mean quality of each of the ten options was chosen

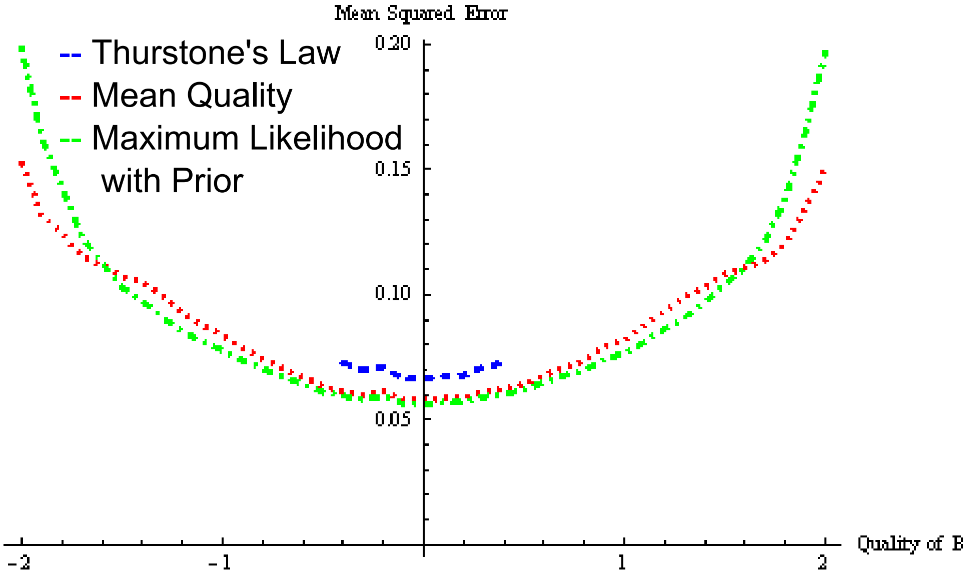

Figure 6: Result of 10,000 trials with two options. In each trial 25 people judge the paired comparison. The mean quality of option A is fixed at 0, and the mean quality of option B is varied along the x axis.

uniformly on $[-x, x]$.

Fig. 7 compares the Bradley-Terry model (labelled "BTL") with the Thurstone model (all results not labelled "BTL"). We also compare the different methods of solving the 0/1 problem (Section 3.2): the least squares methods ("LS") where any 0/1 proportions were"fixed" according to (12) (denoted by "0/1 fix"), Morrissey and Gulliksen's incomplete matrix solution [8, 4] (denoted by "incomplete"), maximum likelihood ("ML"), and the maximum a posteriori estimate where the prior is independently and identically a standard normal on each of the quality scores ("MAP", from (16)). We show two different metrics in Fig. 7:
Interval Mean Squared Error: the average squared error in the quality scale difference for each option pair.

$$\begin{aligned} S_{i,j} &= \mu_i - \mu_j = \text{ ground truth quality difference} \\ S^*_{i,j} &= \mu^*_i - \mu^*_j = \text{ estimated quality difference} \\ \text{Interval MSE} &= \sum_{i \neq j} \frac{(S_{i,j} - S^*_{i,j})^2}{m(m-1)}. \end{aligned}$$

Probability Mean Squared Error: the average squared error in the estimated choice probability for each option pair.

$$\begin{aligned} P_{i,j} &= P(Q_i > Q_j) = \Phi(\mu_i - \mu_j) = \text{ ground truth choice probability} \\ P^*_{i,j} &= \Phi(\mu^*_i - \mu^*_j) = \text{ estimated choice probability} \\ \text{Choice Probability MSE} &= \sum_{i \neq j} \frac{(P_{i,j} - P^*_{i,j})^2}{m(m-1)}. \end{aligned}$$

When the true qualities are close together, the BTL logistic model performs slightly better than the Thurstone-Mosteller Gaussian model because the logistic CDF has a steeper slope when the probability is $\frac{1}{2}$ so it estimates slightly lower values. The least squares methods perform worse as the true means become more separated, but the maximum likelihood methods perform better.

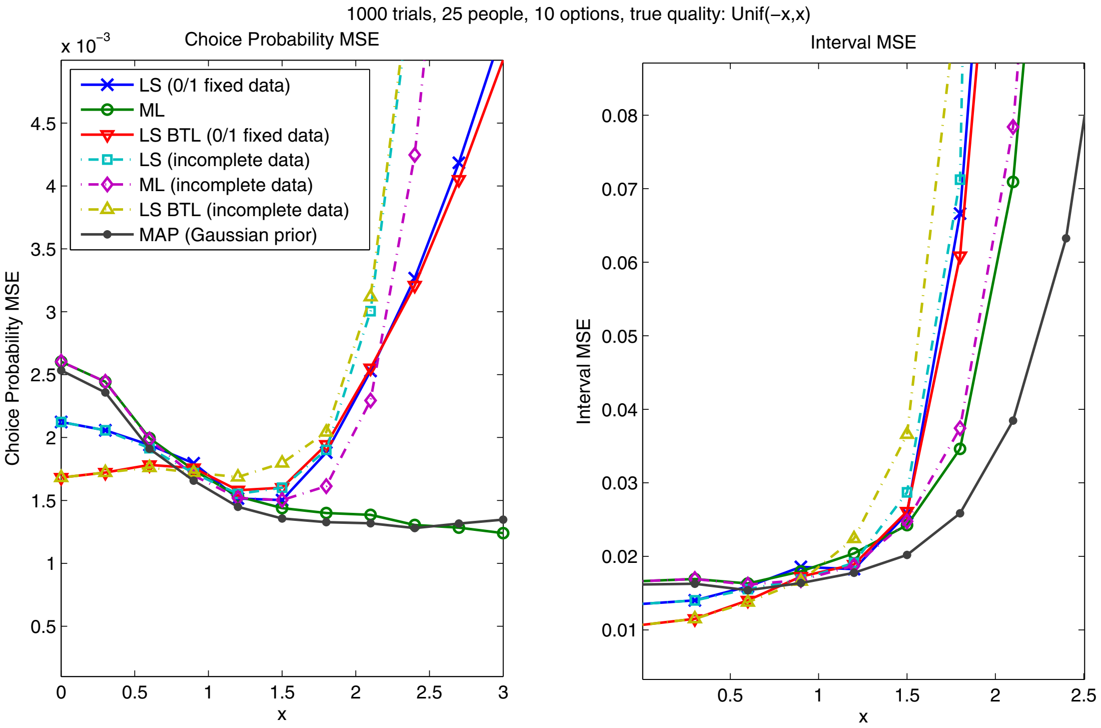

Figure 7: Result of 1,000 trials comparing ten options. In each trial 25 people judge each paired comparison.

7 Summary

Fitting Thurstone's model or the BTL model to paired comparison data can be a useful tool to analyze the relative qualities of a set of options. Various estimation methods can be used to fit each model. If the true quality differences are believed to be separated by at least a standard deviation, then we suggest using the maximum a posteriori estimate is advantageous because we have seen it consistently produce good results, it is an optimal approach to estimation, the data does not need to be modified to avoid 0/1 problems, and it can be solved efficiently as a convex optimization problem.

8 Appendix

The maximum likelihood solution (15) is equivalent to solving for the $\mu_{ij} = \mu_i - \mu_j$ that solves

$$\begin{aligned} \underset{\forall \mu_{ij}}{\arg\max} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_{ij})) \\ \text{subject to} \quad & \mu_{ij} + \mu_{jk} = \mu_{ik} \quad \forall\, i,j,k \in \{1, ..., m\} \end{aligned}$$

The likelihood $\sum_{i,j} C_{i,j} \log(\Phi(\mu_{ij}))$ is concave if Φ is a log-concave function, since $\log(\Phi(x))$ would be concave and concavity is preserved under addition and positive scaling.

Definition. *Log-concave. A function $f : \mathbb{R}^n \to \mathbb{R}$ is log-concave if $f(x) > 0$ and $g(x) = \log(f(x))$ is concave (or equivalently if $h(x) = -\log(f(x))$ is convex).*

Proof. Use the fact that log-concavity is closed under convolution. Since the Gaussian PDF is log-concave (the second derivative of its log is $\frac{d^2}{dx^2} \log \phi(x) = -1$), and the CDF is the convolution of the Gaussian PDF with the unit step function (which is also log-concave). □

The concavity of the Gaussian CDF may alternatively be proved because the CDF of a differentiable log-concave probability density is log-concave. The Gaussian is a differentiable log-concave density function, so the CDF of a Gaussian is log-concave.

Corollary 1. *$f(x)$ is log-concave iff $f(x) > 0$ and $(f'(x))^2 \geq f''(x)f(x)$*

This follows from the condition that $h(x) = -\log(f(x))$ is convex iff the second derivative of $h(x)$ is greater than or equal to zero.

Lemma 1. *The cumulative distribution function of a log-concave differentiable probability density is log-concave.*[4]

Proof. Let $g(t) = \exp(-h(t))$ be a differentiable log-concave probability density function and let the cumulative distribution be

$$f(x) = \int_{-\infty}^{x} g(t)\, dt = \int_{-\infty}^{x} e^{-h(t)}\, dt.$$

We prove that f is log-concave by showing that f satisfies corollary 1.

The derivatives are

$$\begin{aligned} f'(x) &= g(x) = e^{-h(x)} \\ f''(x) &= g'(x) = -h'(x)e^{-h(x)} = -h'(x)g(x). \end{aligned}$$

For $h'(x) \geq 0$, since $f(x) > 0$ and $g(x) > 0$,

$$\begin{aligned} f(x)f''(x) &= -f(x)h'(x)g(x) \leq 0 \\ (f'(x))^2 &= g(x)^2 > 0. \end{aligned}$$

Therefore, Corollary 1 holds.

To show Corollary 1 holds for $h'(x) < 0$, we note that since h is convex,

$$h(t) \geq h(x) + h'(x)(t - x).$$

[4] This lemma appears in Boyd and Vandenberghe's Convex Optimization [1], as exercise 3.55.

Taking the negative, exponent and integrating both sides,

$$\begin{aligned}\int_{-\infty}^{x} e^{-h(t)}\,dt &\leq \int_{-\infty}^{x} e^{-h(x)-h'(x)(t-x)}\,dt\\ &= e^{-h(x)+xh'(x)}\int_{-\infty}^{x} e^{-th'(x)}\,dt\\ &= e^{-h(x)+xh'(x)}\frac{e^{-xh'(x)}}{-h'(x)}\\ &= \frac{e^{-h(x)}}{-h'(x)}.\end{aligned}$$

Multiplying both sides by $-h'(x)e^{-h(x)}$,

$$\begin{gathered}-h'(x)e^{-h(x)}\int_{-\infty}^{x} e^{-h(t)}\,dt \leq e^{-2h(x)}\\ f''(x)f(x) \leq (f'(x))^2.\end{gathered}$$

□

9 Code

```
function q = thurstone(a,b)
% This function returns the estimate in the difference of the means
% according to Thurstone's law of comparative judgment (Case V).
% Each quality distribution is assumed to have variance = 1/2
% so that the variance of the difference of any two quality
% random variables is 1.
%
% Usage:
% q = thurstone(a,b)
% Input integers a and b representing the number of times
% object a or b was preferred.
% Example: q = thurstone(2, 3);
%
% q = thurstone(v)
% Input vector v is the counts for the number of times that
% v(1) or v(2) was preferred.  Only supported for vectors of length 2.
% Example: q = thurstone([2 3]);
% This gives the same result as the example for thurstone(2, 3)
% above.
%
% q = thurstone(p)
% Input can also be a scalar, p for the proportion of times
% that C1 was chosen over C2.
% Example: q = thurstone(0.4);
% This gives the same result as the example for vector C above.
%
% q is the estimated difference in the means of C(1) - C(2)
%
% S = thurstone(C)
% C is a square matrix of counts, where
%   C(i,j) = # times option i was preferred over option j
% and it is assumed that
%   C(i,j) + C(j,i) = number of times that option i was judged against j.
% Note that diag(q) = 0 since comparing any 2 same options
% should result in a 50-50% chance of chosing each option.
% Example: S = thurstone([0 3; 2 0]);
% S(1,2) should be equal to the above examples.
% S(2,1) should be equal to the negative of the above examples.
% Generally S == −S'.

%=======
% Kristi Tsukida <kristi.tsukida@gmail.com>
% Created Jan 11, 2010
% Last Updated Mar 3, 2010
% added matrix input
%=======

% Input checks
if nargin > 1
    assert(isnumeric(a), 'a must be numeric');
    assert(isnumeric(b), 'b must be numeric');
    assert(isscalar(a), 'a must be a scalar');
    assert(isscalar(b), 'b must be a scalar');
    assert(¬isnan(a), 'a cannot be NaN');
    assert(¬isnan(b), 'b cannot be NaN');
    assert(a>0, 'a must be positive');
    assert(b>0, 'b must be positive');

    p=a/(a+b);
elseif isscalar(a)
    p=a;
    assert(isnumeric(p), 'p must be numeric');
    assert(¬isnan(p), 'p cannot be NaN');
    assert(p≥0, 'p must be positive');
    assert(p≤1, 'p must be less than 1');
elseif isvector(a)
    v=a;
    %assert(length(v)==2, 'function not defined for vector v longer than 2');
    assert(isnumeric(v), 'v must be numeric');
    assert(all(v≥0), 'counts in v must be positive');
    assert(sum(v)>0, 'counts in v must have positive sum');
    % vector v
    if(length(v)==2)
        p = v(1) ./ sum(v);
    else
        assert(all(v≤1), 'probabilities in v must be less than 1');
        p = v;
    end

else % matrix input
    C=a;
    assert(isnumeric(C), 'C must be numeric');
    assert(all(C(:)≥0), 'counts in C must be positive');
    assert(sum(C(:))>0, 'counts in C must have positive sum');
    assert(size(C,1)==size(C,2), 'C must be a square matrix');

    N = C + C';
    % matrix of probabilities
    % (norminv handles matrices of probabilities)
    p = C ./ N;
    % fix p so that q ends up with zeros on the diagonal.
    p(speye(size(p))>0) = 0.5;
end

% Assuming independence, and each quality RV has variance = 1/2,
% the variance of C1 − C2 is 1
sigma = 1;
% Calculate q
q = norminv(p, 0, sigma);
% same as q = norminv(p, 0, 1)*sigma;
% norminv(p, 0, 1) is the zscore of p
```

```
function S = scale_ls(counts, threshold)
% Use the least squares complete matrix solution
% (Thurstone, Mosteller) to
% scale a paired comparison experiment using
% Thurstone's case V model (assuming sigma^2 = 0.5 for each
% quality's distribution)
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values

if nargin < 2
    threshold = 2;
end

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
P = counts ./ (counts + counts');
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

Z = norminv(P);
S = sum(Z,1)' / m;
```

```
function S = scale_gulliksen(counts, threshold)
% Use the Morrisey-Gulliksen incomplete matrix solution to
% scale a paired comparison experiment using
% Thurstone's case V model (assuming sigma^2 = 0.5 for each
% quality's distribution)
%
% (This code follows Gulliksen's formulation given in Engeldrum's
% Psychometric Scaling book)
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values
%   Scale values are set up to have mean of 0

if nargin < 2
    % default threshold on scale difference
    threshold = 2;
end

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
P = counts ./ (counts + counts');
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

% Thurstone's law estimates of each pairwise quality difference
% (norminv calculates the z-scores or z-value)
% ! entries in Z could end up NaN
Z = norminv(P);

M = double(abs(Z) > threshold); % 1 where |Z(i,j)| > 2, 0 otherwise
M(eye(m)>0) = m - sum(M); % set the diagonal values

d = sum(Z, 2); % = # of valid comparisons + 1

S = M \ d; % = inv(M) * d;
```

```
function S = scale_ml(counts, threshold)
% Use cvx to compute maximum likelihood scale values

if nargin < 2
    % default threshold on scale difference
    threshold = 2;
end

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

previous_quiet = cvx_quiet(1);
cvx_begin
    variables S(m,1) t;
    U = repmat(S,1,m);
    △ = U − U'; % △(i,j) = S(i) − S(j)

    minimize( t );
    subject to
    −sum(sum(counts.*log_normcdf(△))) ≤ t
    sum(S)==0
cvx_end
cvx_quiet(previous_quiet);
```

Acknowledgements

This work was supported by the United States Office of Naval Research and the United States PECASE award.

References

[1] S. P. Boyd and L. Vandenberghe. Convex optimization, 2010.

[2] R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324, December 1952.

[3] P. G. Engeldrum. Psychometric scaling : a toolkit for imaging systems development, 2000.

[4] H. Gulliksen. A least squares solution for paired comparisons with incomplete data. *Psychometrika*, 21:125–134, 1956.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[6] R. D. Luce. Individual choice behavior; a theoretical analysis., 1959.

[7] R. D. Luce. Thurstone and sensory scaling: Then and now. *Psychological Review*, 101(2):271–277, 1994.

[8] J. H. Morrissey. New method for the assignment of psychometric scale values from incomplete paired comparisons. *Journal of the Optical Society of America*, 45(5):373–8, May 1955.

[9] F. Mosteller. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, March 1951.

[10] S. S. Stevens. On the Theory of Scales of Measurement. *Science (New York, N.Y.)*, 103(2684):677–80, June 1946.

[11] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 1927.